Dr. Nan Tang

Senior Scientist, Data Analytics, QCRI
Room1814, Tornado tower, West Bay
Doha, Qatar

📱 *+974 66700540*
☎ *+974 44542850*
✉ ntang@qf.org.qa
http://web.qcri.org/~ntang/

Education

2004–2007 **Ph.D.**, Systems Engineering & Engineering Management, *The Chinese University of Hong Kong*. Thesis: Efficient XPath Query Processing in Native XML Databases (co-supervisors: Prof. Jeffrey Xu Yu and Prof. Kam-Fai Wong).

2001–2004 **M.E.**, Computer Science, *Northeastern University*, China. Thesis: Parallel XML Databases.

1997–2001 **B.E.**, Computer Science, *Northeastern University*, China.

Experience

04/2015-now *Senior Scientist*, Data Analytics, **Qatar Computing Research Institute, HBKU**, Qatar.

12/2011-03/2015 *Scientist*, Data Analytics, **Qatar Computing Research Institute, HBKU**, Qatar.

02/2010-01/2012 *Research fellow*, **University of Edinburgh**, UK. Worked on data cleaning and social graphs, with Prof. Wenfei Fan.

02/2008-01/2010 *Scientific staff member*, **CWI** (Dutch National Research Center for Mathematics and Computer Science), the Netherlands. Worked on column-store database MonetDB and distributed XQuery processing, with Dr. Peter Boncz.

03–08/2007 *Visiting Scholar*, **University of Waterloo**, Canada. Worked on XML indexing & query rewriting, with Prof. Tamer Özsu.

Awards and Honors

2014 Asia-Pacific Web Conference (APWeb) 2014, ***Distinguished Lecturer***.

2012 Selected for **best papers of ICDE 2012**, for my paper, titled "Incremental Detection of Inconsistencies in Distributed Data".

2010 The 37th VLDB conference, **the Best Paper award of VLDB 2010**, for my paper, titled "Towards Certain Fixes with Editing Rules and Master Data".

2009 Selected for **best papers of ICDE 2009**, for my paper, titled "Projective Distribution of Full-Fledged XQuery".

Research Interests Projects

Broad Interests: Data Curation, Network Monitoring, and Big Data Analytics.

Data curation
2012-today

• NADEEF: a commodity data cleaning system that is extensible, generic and easy-to-deploy. Nan is the PI of this project. It made the following achievements:

- ○ an open-source data cleaning system
- ○ a portfolio of $>$ 10 applied patents, two research papers, and two system demos
- ○ under commercialization in Qatar Science & Technology Park

• KATARA: using knowledge bases and crowdsourcing for reliable data cleaning. Nan is the PI of this project.

• PEARL: an ensemble method using active learning and transfer learning to detect data errors with accuracy guarantees. Nan is a co-PI of this project.

• Interactive data cleaning. In data cleaning, for some applications, involving users is necessary to improve data quality. This direction resulted in two independent collaborative projects:

◦ QCRI-MIT. It applies the tool *program by synthesis* for data cleaning via interacting with users. Nan is the PI from QCRI side. Prof. Sam Madden and Prof. Armando Solar are my collaborators from MIT.

◦ QCRI-Tsinghua: We use SQL-update queries as the rules to update (clean) the data. We try to infer such rules by minimizing user efforts. Nan is the PI from QCRI. Prof. Guoliang Li is my collaborator from Tsinghua University.

Network monitoring
2015-today

Monitoring network data has wide applications for cyber security and social networks. Nan is the PI of this project. I am working on two directions.

◦ network data summarization, or more generally, sketches for high-dimension data streams

◦ mining/prediction over historical sketches

Big Data Analytics
2013-today

Rheem: big data analytics over diverse data processing platforms. We target at achieving platform independence and performance benefits for various domains, such as data cleaning, machine learning and graph management. Nan is a co-PI of this project.

Closed projects

Column-stores

• *Update module*: implemented packed-memory array for managing sorted columns in MonetDB, which improves per update cost from $O(N)$ to $O(log^2 N)$ amortized element moves.

String matching

• Designed effective partial posting list selection algorithm, and implemented bitwise data compression, which saves *87.5%* storage overhead on huge strings (up to *426GB*).

• Implemented scalable algorithms for huge disk-based strings, which use a two-pass external sorting and a cache-conscious radix-sort. This improves the speed by *an order of magnitude*.

XML databases

• *XIRAF*: XML-based indexing and querying for digital forensics, a project developed for national forensic centre of the Netherlands. My role was to rewrite XPath queries using materialized XPath views, which saved query evaluation time up to *90%*.

• *XRPC*: designed a simple XQuery extension that allows efficient and interoperable distributed queries for full-fledged XQuery (including update facility), using SOAP requests.

Social graphs

• Proposed novel graph pattern queries, which can be evaluated in cubic time. Notably, its traditional counterpart that utilizes subgraph isomorphism is NP-complete.

• Incorporated regular expressions to meet the requirements of emerging applications.

Teaching and Mentoring Experience

Internship Mentor, Qatar Computing Research Institute

Qing Chen, Master student from Fudan University, China.
- Project: Network data summarization. 2015/07-now

Shuai Li, Ph.D. student from University of Insubria, Italy.
- Project: Mining/prediction over network data streams. 2015/08-now

Jian He, Master student from Tsinghua University, China.
- Project: Falcon: Interactive data cleaning. 2014/11-2015/02

Matteo Interlandi, Ph.D. student from University of Modena, Italy.
- Project: declarative data annotation. 2014/03-2014/05

Chu Xu, Ph.D. student from University of Waterloo, Canada.
- Project: Katara: cleaning with knowledge bases and crowdsourcing. 2013/05-2014/07

Ahmed Eldawy, Ph.D. student from University of Minnesota, US.
- Project: Nadeef: generalized data cleaning. 2012/01-2012/05

Michele Dallachiesa, Ph.D. student from University of Trento, Italy.
- Project: Nadeef: generalized data cleaning. 2012/01-2012/05

Yu Tang, Master student from Hong Kong University, Hong Kong.
- Project: Nadeef dashboard. 2012/11-2013/01

Jiannan Wang, Ph.D. student from Tsinghua University, China.
- Project: automated and dependable data repairing. 2012/12-2013/02

Amr Ebaid, Ph.D. student from Purdue University, US.
- Project: Nadeef: generalized data cleaning. 2012/04-2013/01

Teaching, University of Edinburgh (Tutorials)

Applied Databases, 09-11/2010

Teaching, The Chinese University of Hong Kong (Tutorials)

Fundamentals of Information Systems, 09-12/2004, 01-05/2006
Information Systems Design & Analysis, 01-05/2005
Digital Logical and Systems, 09-12/2006, 09-12/2007

Professional Activities and Services

Book Chapter reviewer — Advanced Applications and Structures in XML Processing: Label Streams, Semantics Utilization and Data Query Technologies.

Journal reviewer — VLDB Journal, 2009, 2010, 2011
ACM Transactions on Knowledge Discovery from Data (TKDD), 2012
ACM Transactions on Database Systems (TODS), 2013
ACM Transactions on the Web (TWEB), 2012
IEEE Trans. on Knowledge and Data Engineering (TKDE), 2007,2011,2012
World Wide Web Journal, 2009, 2011
Data and Knowledge Engineering (DKE), 2010
ACM Computing Reviews, 2010
Knowledge and Information Systems, 2009

PC-Member — International Conference on Very Large Data Bases (PVLDB), 2015
ACM SIGMOD International Conference on Management of Data (SIGMOD), 2015
ACM Conference on Information and Knowledge Management (CIKM) 2011, 2012
IEEE International Conference on Data Engineering (ICDE), 2013
International Workshop on Graph-structured Data Bases 2011
Asia Pacific Web Conference 2008, 2010
The Joint International Conferences on APWeb and WAIM 2009

Publications

Journal Publications

1. Zuhair Khayyat, William Lucia, Meghna Singh, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, Nan Tang and Panos Kalnis. *Lightning Fast and Space Efficient Inequality Joins*. PVLDB, 2015. (The extended version has been *invited* to VLDB journal, which is under preparation.)
2. Wenfei Fan, Floris Geerts, Nan Tang, and Wenyuan Yu. *Conflict Resolution with Data Currency and Consistency*. ACM Journal of Data and Information Quality (JDIQ), 2014 (*invited*).

3. Wenfei Fan, Shuai Ma, Nan Tang, and Wenyuan Yu. *Interaction between Record Matching and Data Repairing.* ACM Journal of Data and Information Quality (JDIQ), 2014.
4. George Beskales, Gautam Das, Ahmed K. Elmagarmid, Ihab F. Ilyas, Felix Naumann, Mourad Ouzzani, Paolo Papotti, Jorge Quiane-Ruiz, and Nan Tang. *The Data Analytics Group at the Qatar Computing Research Institute.* SIGMOD Record, 2012.
5. Wenfei Fan, Jianzhong Li, Nan Tang, and Wenyuan Yu. *Incremental Detection of Inconsistencies in Distributed Data.* IEEE Transaction on Knowledge and Data Engineering (TKDE), 2014 (Special issue: *Best Papers of ICDE 2012, invited*).
6. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Yinghui Wu. *Adding Regular Expressions to Graph Reachability and Pattern Queries.* Frontiers of Computer Science, 2012 (Special issues: New Topics in Database, *invited*).
7. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Wenyuan Yu. *Towards Certain Fixes with Editing Rules and Master Data.* VLDB Journal, 2012 (Special issue: *Best Papers of VLDB 2010, invited*).
8. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Wenyuan Yu. *Towards Certain Fixes with Editing Rules and Master Data.* In PVLDB, 2010 (*The Best Paper award*).
9. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Yinghui Wu, Yunpeng Wu. *Graph Pattern Matching: From Intractable to Polynomial Time.* In PVLDB, 2010.
10. Ying Zhang, Nan Tang, Peter Boncz. *Projective Distribution of Full-Fledged XQuery.* IEEE Transaction on Knowledge and Data Engineering (TKDE), 2010 (Special issue: *Best Papers of ICDE 2009, invited*).
11. Nan Tang, Jeffrey Xu Yu, Kam-Fai Wong, and Jianxin Li. *Fast XML structural join algorithms by partitioning.* Journal of Research and Practice in Information Technology (JRPIT), 2008.
12. Kam-Fai Wong, Jeffrey Xu Yu, and Nan Tang. *Answering XML queries using path-based indexes: A survey.* World Wide Web Journal, 2006.

Conference Publications

1. Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang and Yin Ye. *KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing.* SIGMOD, 2015.
2. Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang and Yin Ye. *KATARA: Reliable Data Cleaning with Knowledge Bases and Crowdsourcing.* VLDB demo, 2015.
3. Zuhair Khayyat, Ihab F. Ilyas, Alekh Jindal, Sam Madden, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiane-Ruiz, Nan Tang and Si Yin. *BigDansing: A System for Big Data Cleansing.* SIGMOD, 2015.
4. Nan Tang. *Big RDF Data Cleaning.* The 6th International Workshop on Data Engineering meets the Semantic Web (DESWeb), in conjunction with ICDE, 2015 (*invited*).
5. Matteo Interlandi and Nan Tang. *Proof Positive and Negative in Data Cleaning.* ICDE, 2015.
6. Nan Tang. *Big Data Cleaning.* APWeb, 2014 (*invited as Distinguished Lecture Series*).
7. Jiannan Wang and Nan Tang. *Towards Dependable Data Repairing with Fixing Rules.* SIGMOD, 2014.
8. Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge Quiane-Ruiz, Nan Tang, and Si Yin. *NADEEF/ER: Generic and Interactive Entity Resolution.* SIGMOD demo, 2014.
9. Wenfei Fan, Floris Geerts, Shuai Ma, Nan Tang, and Wenyuan Yu. *Data Quality Problems beyond Consistency and Deduplication.* In search of elegance in the theory and practice of computation: a Festschrift in honour of Peter Buneman, Edinburgh, UK, 2013 (*invited*).

10. Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and Nan Tang. *NADEEF: A Commodity Data Cleaning System.* SIGMOD, 2013.
11. Amr Ebaid, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge Quiane-Ruiz, Nan Tang, and Si Yin. *NADEEF: A Generalized Data Cleaning System.* VLDB demo, 2013.
12. Wenfei Fan, Floris Geerts, Nan Tang, and Wenyuan Yu. *Inferring Data Currency and Consistency for Conflict Resolution.* ICDE, 2013.
13. Wenfei Fan, Jianzhong Li, Nan Tang, and Wenyuan Yu. *Incremental Detection of Inconsistencies in Distributed Data.* ICDE, 2012.
14. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Wenyuan Yu. *CerFix: A System for Cleaning Data with Certain Fixes.* VLDB demo, 2011.
15. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Wenyuan Yu. *Interaction between record matching and data repairing.* SIGMOD, 2011.
16. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Yinghui Wu. *Adding Regular Expressions to Graph Reachability and Pattern Queries.* ICDE, 2011.
17. Nan Tang, Lefteris Sidirourgos, Peter Boncz. *Space-Economical Q-Gram Index for Exact String Matching.* CIKM, 2009.
18. Ying Zhang, Nan Tang, Peter Boncz. *Efficient Distribution of Full-Fledged XQuery.* ICDE, 2009.
19. Nan Tang, Jeffrey Xu Yu, Hao Tang, M. Tamer Özsu and Peter Boncz. *Materialized View Selection in XML Databases.* DASFAA, 2009.
20. Nan Tang, Jeffrey Xu Yu, M. Tamer Özsu, Byron Choi, and Kam-Fai Wong. *Multiple materialized view selection for XPath query rewriting.* ICDE, 2008.
21. Nan Tang, Jeffrey Xu Yu, M. Tamer Özsu, and Kam-Fai Wong. *Hierarchical indexing approaches to support XPath queries.* ICDE, 2008.
22. Nan Tang, Jeffrey Xu Yu, Kam-Fai Wong, and Haifeng Jiang. *Fast structural join with a location function.* DASFAA, 2006.
23. Jiefeng Cheng, Jeffrey Xu Yu, and Nan Tang. *Fast reachability query processing.* DASFAA, 2006.
24. Nan Tang, Jeffrey Xu Yu, Kam-Fai Wong, Kevin Lü, and Jianxin Li. *Accelerating XML structural join by partitioning.* In Proc. 16th International Conference on Database and Expert Systems Applications (DEXA), 2005.
25. Nan Tang, Guoren Wang, Jeffrey Xu Yu, Kam-Fai Wong, and Ge Yu. *Win: An effcient data placement strategy for parallel XML databases.* In Proc. 11th International Conference on Parallel and Distributed System (ICPADS), 2005.
26. Bing Sun, Bo Zhou, Nan Tang, Guoren Wang, Ge Yu, and Fulin Jia. *Answering XML twig queries with automata.* In Proc. 6th Asia-Pacific Web Conference (APWeb), 2004.
27. Yaxin Yu, Guoren Wang, Ge Yu, Gang Wu, Junan Hu, and Nan Tang. *Data placement and query processing based on RPE parallelisms.* In Proc. 27th International Computer Software and Applications Conference (COMPSAC), 2003.